Technical Disclosure Commons

Defensive Publications Series

August 2020

Topic Granularity Detection Based On Historical Distribution

Anonymous

Topic Granularity Detection Based On Historical Distribution

ABSTRACT

User generated hashtags are widely used on social media platforms, blogs, and community sites to annotate content and make it easily discoverable by humans and algorithms. Hashtags reflect the semantics of the content from the users' points of view, in their own vocabulary. Detecting the relationship between hashtags is often a challenging problem. This disclosure describes techniques that leverage the historical distribution of topics in social media posts or other content to accurately detect candidates for higher level topics in a topical graph. A generalization score is calculated based on the observation that higher level topics typically have a long history with a nearly even distribution. Topics that have a longer history and a consistent presence across various time periods get a higher generalization score and can be accurately detected as having coarser granularity.

KEYWORDS

- Hashtag
- Social media
- Information sharing
- Information retrieval
- Topic recommendation
- Topic suggestion
- Hierarchical graph
- Hashtag Abuse

BACKGROUND

User generated hashtags are used widely on social media platforms, blogs, and community sites to annotate content and make it easily discoverable by humans and algorithms. Hashtags play a key role in facilitating information sharing and information retrieval. Hashtags reflect the semantics of the content from the users' points of view, in their own vocabulary. Since hashtags represent the users' point of view about the content at a point in time, there is a wide variety of hashtags that represent the same or similar content across time and various user groups.

While there may be a hierarchy or relationship in the underlying content, it is not represented in the hashtags. Detecting the relationship between hashtags is often a challenging problem. For example, social media posts about football tournament hosted in Boston in 2018 may be associated with a variety of hashtags such as #football, #collegefootball2018, #football2018, #tournament2018, #BostonCollegefootball2018.

In the football season, there may be a substantial number of posts that report the tournament and a number of hashtags describing the football season, or a specific team or player's performance during the season. These hashtags are appropriate as a topic for all the posts related to college football in 2018. However, the same hashtags may not be a good candidate for a higher level topic. For example, these hashtags are not suitable to be used for posts related to the college football season in 2019. Further, these topics would likely be too specific to build a topical graph. Higher level topics such as #collegefootball, #tournament are better candidates for building the topical graph, while #collegefootball2018 is a good candidate as a subtopic.

However, higher level topics (such as #collegefootball) might not appear frequently in posts, as compared to a lower level topic such as (#collegefootball2018). Conventional metrics such as co-concurrence, number of mentions, number of likes are not particularly valuable helpful in detecting #collegefootball as a suitable higher level topic.

DESCRIPTION

This disclosure describes an approach that leverages the historical distribution of topics to accurately detect candidates for higher level topics in a topical graph. The approach is based on the observation that higher level topics typically have a long history with a nearly even distribution. On the other hand, topics that appear only in the recent past are more likely to be lower level, concrete topics.

Typically, hashtags (or topics) are created either directly by the end user or by auto-generation programs from external systems. A post may contain different topics, and a topic may also show up in different posts. The approach described in this disclosure first performs detection of relationships between the various topics. This is achieved through creation of a post-topic mapping graph. The post-topic mapping graph is used to generate topic candidates which are similar or have a “contains” or “belong” relationship. For example, the topics #collegefootball2018 and #football2018 have a contains relationship, since #football2018 is entirely contained in the topic #collegefootball2018. On the other hand, the topics #collegefootball2018, #ncaaplayoffs2018, and #ncaatournament2018 are similar.

A next step is to detect the hierarchy among hashtags that relate to (are descriptive of) the same topic but at a different granularity. This step involves creating a generalization score using the historical distribution of topics to represent topic granularity. The historical distribution of topics can be created using date and time information from the posts. To compute the

generalization score, the posts are split into different buckets based on a time window, such as day, week, month etc.

Topic generalization score considers the following aspects of the historical topic distribution:

- **Vertical distribution:** Probability of each topic within a time bucket. This is a measure of the popularity of a topic among all topics in a given time bucket.
- **Horizontal distribution:** Incidence of a topic in a time bucket as a proportion of its overall incidence in history. This is a measure of the popularity of a topic in a given time bucket as compared to its popularity through history.
- **Entropy of the horizontal distribution:** Measure of the amount of uncertainty in the horizontal distribution. This is a measure of the consistency in the popularity of a topic through history.

The generalization score can be computed based on:

- N - a total number of buckets;
- P(i) - the probability of a particular topic *i* among all posts within a particular bucket;
- P'(i) - the incidence of the topic *i* in the particular bucket as a proportion of its overall incidence; and
- the entropy of the horizontal distribution of topic *i*.

Topics that have a longer and more even distribution across history are assigned a higher generalization score. To detect topic hierarchy, generalization scores of related topics are compared and topics with higher generalization scores are determined to be the coarser grained or higher level topic.

For example, consider the following data: for the topic #collegefootball distribution for P(i) is [1/5, 1/5, 1/5] and horizontal distribution P'(i) is [1/3, 1/3, 1/3]; and for the topic #collegefootball2018, the corresponding distributions are [0,0,1/5] and [0,0,1]. The generalization score for the topic #collegefootball would then be higher than the generalization score for the topic #collegefootball2018, since #collegefootball has a higher entropy due to a consistent horizontal distribution. By comparing the generalization scores, the topic #collegefootball is correctly identified to have coarser granularity compared to the topic #collegefootball2018.

The generalization score computed as described herein can be used in several applications such as topic granularity based recommendation, topic hierarchical graph generation, hashtag abuse detection, etc.

CONCLUSION

This disclosure describes techniques that leverage the historical distribution of topics in social media posts or other content to accurately detect candidates for higher level topics in a topical graph. A generalization score is calculated based on the observation that higher level topics typically have a long history with a nearly even distribution. Topics that have a longer history and a consistent presence across various time periods get a higher generalization score and can be accurately detected as having coarser granularity.